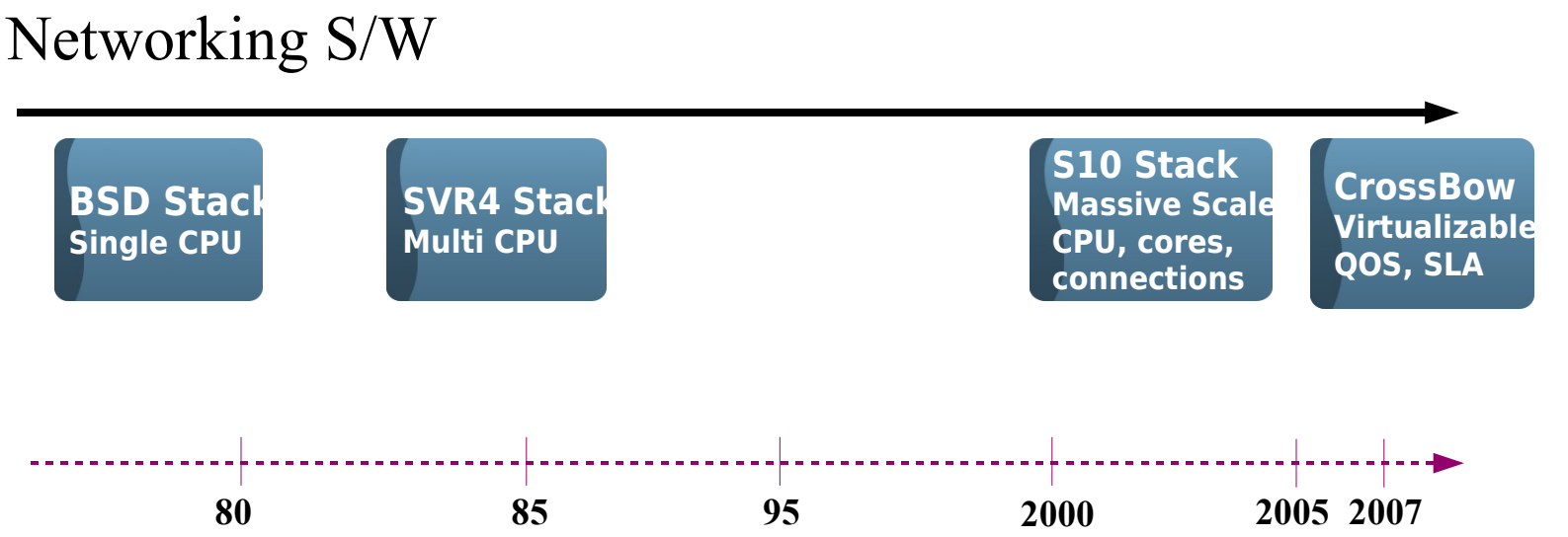
CrossBow:
Network Virtualization

Tutorial

Vineeth Pillai
SUN Microsystems Prague
vineeth.pillai@sun.com

The Network is the Computer



Demand for Virtualization

Financial Services

- Trading house starts offering free financial information to attract customers
- Brokerage customers start complaining that trading site slows down
- The paying customers start deserting

Large ISP

- ISP wants to deploy virtual systems on same physical machines
- ISP sells each virtual system at different price levels to its customers
- Any virtual instance can overwhelmed the shared networking resource

Enterprise Computing

- A large company uses a workgroup server for day to day as well as critical traffic
- IT Ops doing non critical stuff started a backup over the network
- Users doing time critical work can't get bandwidth to do their job

What Happened?

- Critical services are overwhelmed by non-critical services, traffic types, or virtual systems
- No usable mechanism available for fairness, priority and resource control for networking bandwidth

Crossbow: What is it?

- Network virtualization and resource partitioning in OpenSolaris
- Virtual NICs on top of Physical NICs which could be administered as any other physical NIC on the system
- Virtual NIC has its own hardware resources(Rx/Tx rings, DMA channels MAC addresses etc..)
- It virtualizes the IP stack and NIC around any service, protocol, or virtual machine.

Crossbow (contd..)

- Each virtual stack can be assigned its own priority and bandwidth.
- Traffic on one virtual NIC is isolated from traffic on another.
- Maximum utilization of the available network resources in a system.
- No degradation in performance

Crossbow Components

- Crossbow is built on top of
  - Existing Solaris Network Framework(Fire Engine).
  - GLDv3(Generic LAN Driver Framework)
- Crossbow changes the existing architecture and adds new modules to virtualize the Solaris Network Framework and to bring in the resource partitioning.

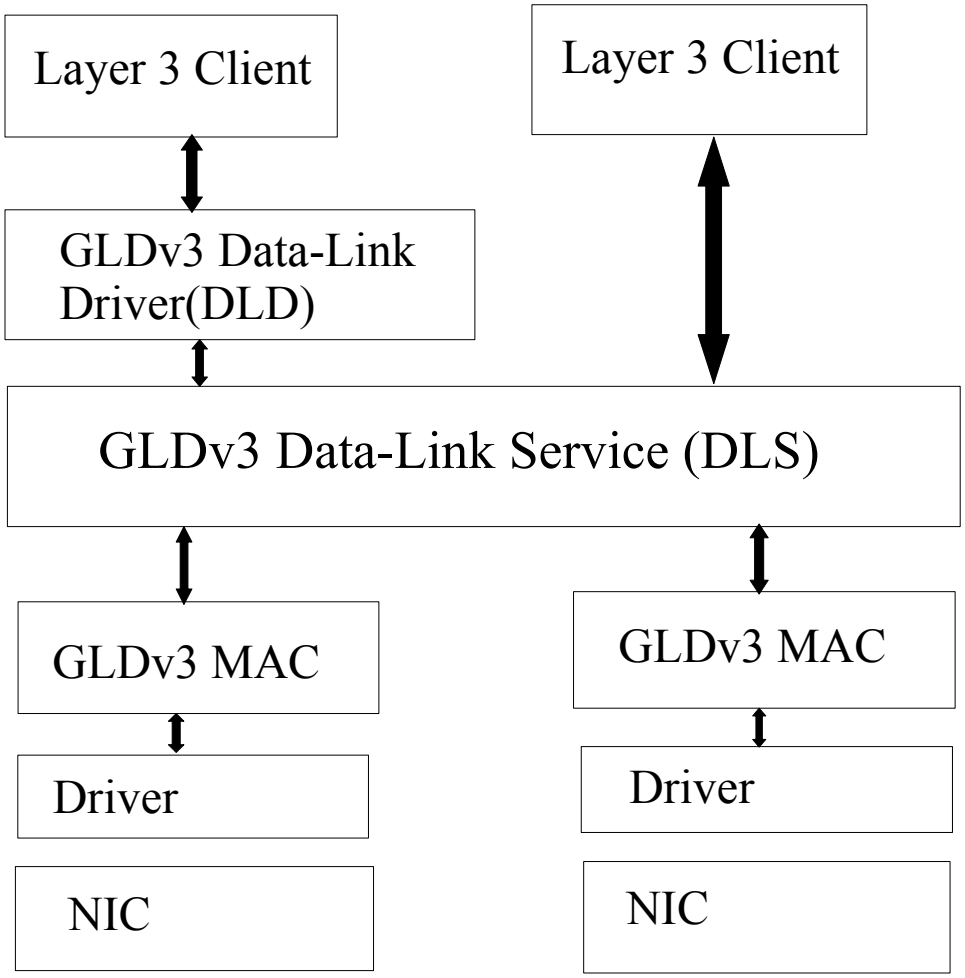
Fire Engine – New network Stack

- Merged IP and TCP Module which is fully multi threaded
- Vertical perimeter : squeues as the synchronization mechanism
  - Per cpu : Better cache locality.
  - Each connection bound to a specific squeue
- BSD style function call interface replaces the current message passing interface.
- STREAMs are still used to provide flexibility that ISVs need to provide additional functionality.

GLDv3(Generic LAN Driver)

- Introduced in Solaris 10 Update1
- Powerful generic LAN interfaces designed to be used by the Network drivers
- Network drivers can focus on managing the hardware resource and leave the rest to GLD.
- Provides capabilities like link aggregation and VLAN features.
- Crossbow built on top of GLDv3.

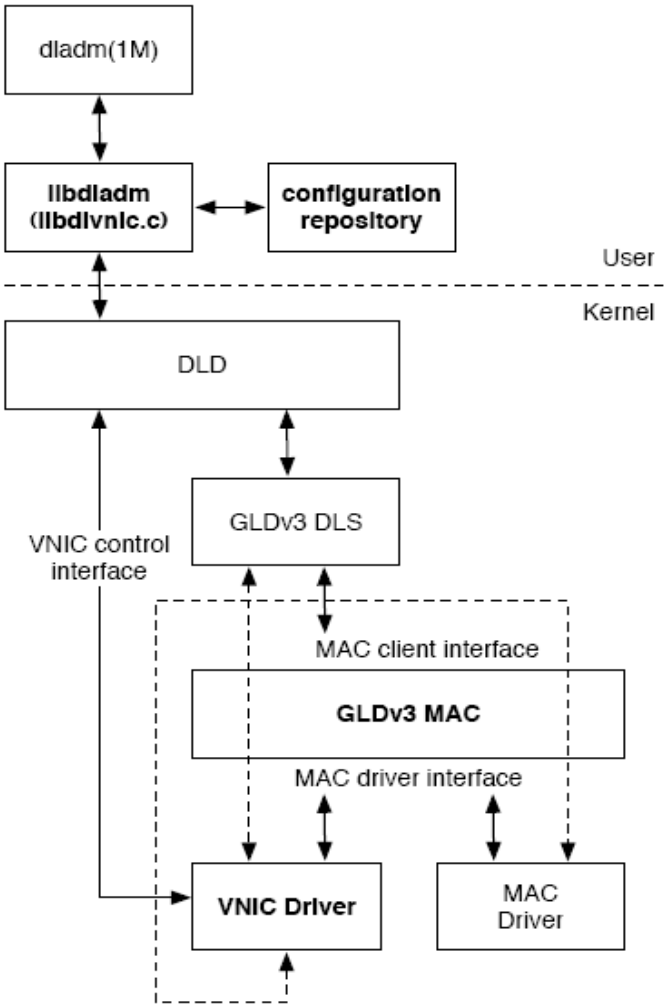
GLDv3 Architecture



Crossbow and GLDv3

- Crossbow changes the MAC layer and introduces VNIC pseudo driver
- MAC layer is modified to allow multiple MAC clients to access the underlying network hardware
- VNIC driver creates a new MAC client which uses the modified MAC client interface of the MAC layer

Crossbow Virtualization Architecture



VNIC Driver

- Pseudo Driver
- Allows the creation of multiple virtual MAC interfaces on top of a single physical NIC or ling aggregation
- At the top, it registers multiple MACs with the MAC layer
- At the bottom, it uses the MAC client interface to access the underlying NIC

Crossbow Features

- Stack and NIC Virtualization
- Resource partitioning, QoS/Diffserv (without performance penalties)
- SLA on a per connection basis
- Better Defense against DDOS attacks
- Real time usage and history
- performance
  - Polling on forwarding path (performance)
  - S/W fanout to multiple cores (utilization)
- Class of service support
- ISV support (APIs for configuration, statistics, traps, etc)
- Network Device Consolidation

The Crossbow Architecture

- Divide NIC memory, DMA channels, etc and use a flow classifier to build a virtual stack on each H/W partition
- Each Virtual NIC is owned by the FireEngine Squeue's which independently switch the VNIC between interrupt & polling mode
- Rate of packet arrival from a VNIC is independently controlled by the Squeue owning the VNIC
- Virtual stack priority is controlled by the squeue thread which does the Rx/Tx processing.

Crossbow: NIC handling

- The architecture supports non Nemo NICs as well as Nemo NICs which don't have flow classification capabilities
- We simulate multiple queues or memory area in the Nemo layer using a S/W flow classifier
- Nemo provides a DLPI shim layer for non Nemo drivers
- All the general 1Gb and 10Gb NICs in future will support the flow classification and memory partitioning capability at no extra cost.

Effects of Dynamic Polling

- Sample mpstat output

Mpstat (older driver)

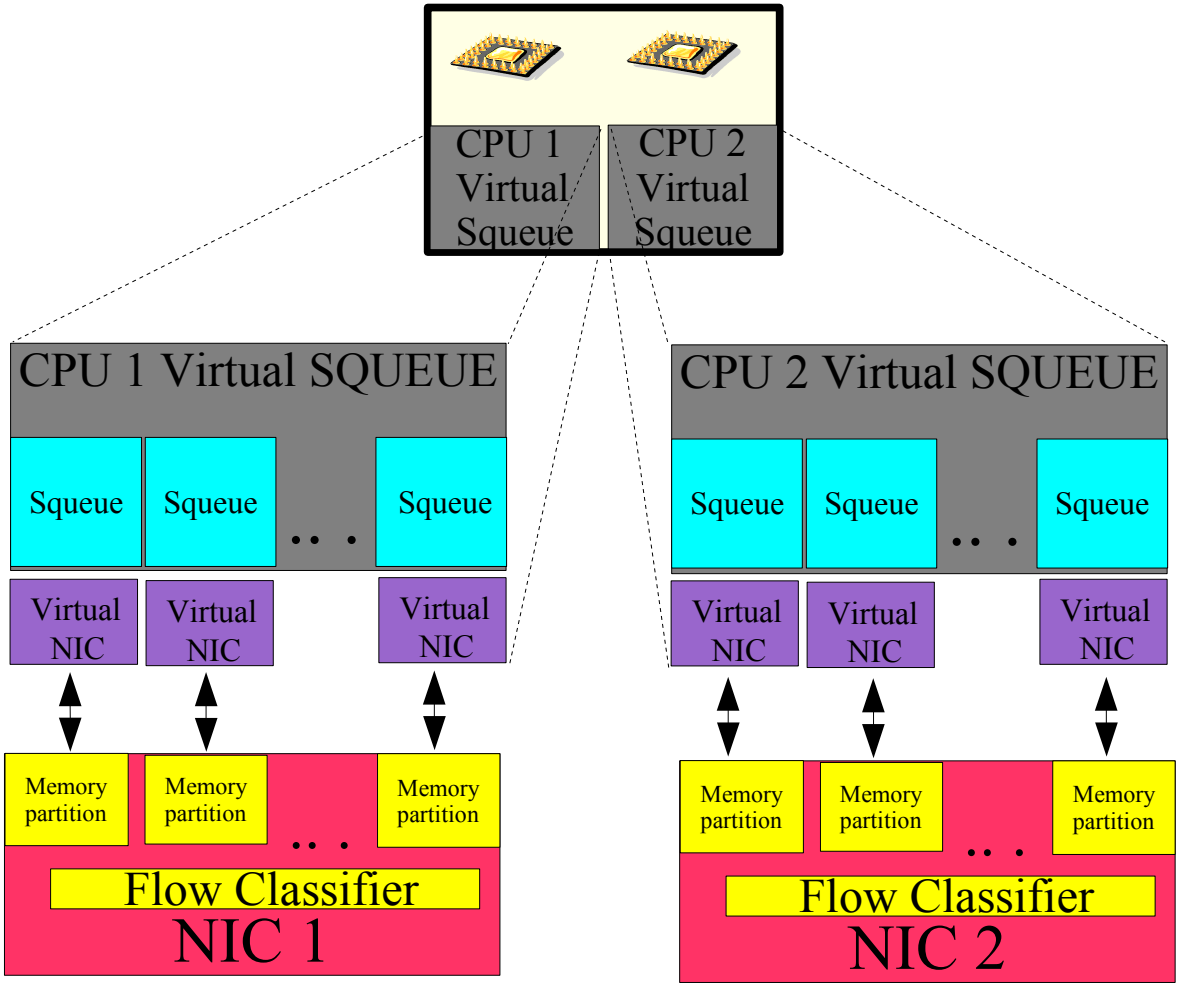
```
intr   ithr  csw   icsw  migr smtx  srw   syscl  usr sys wt idl
10818  8607  4558  1547  161  1797  289   19112  17  69  0  12
```

Mpstat (GLDv3 based driver)

```
intr   ithr  csw   icsw  migr smtx  srw   syscl  usr sys wt idl
2823   1489  875   151   93   261   1     19825  15  57  0  27
```

- Notice the decrease in interrupts, context switches, mutex contentions, etc. and increase in idle time
- Crossbow allows each VNIC to be dynamically polled by its Squeue

Virtual Stacks



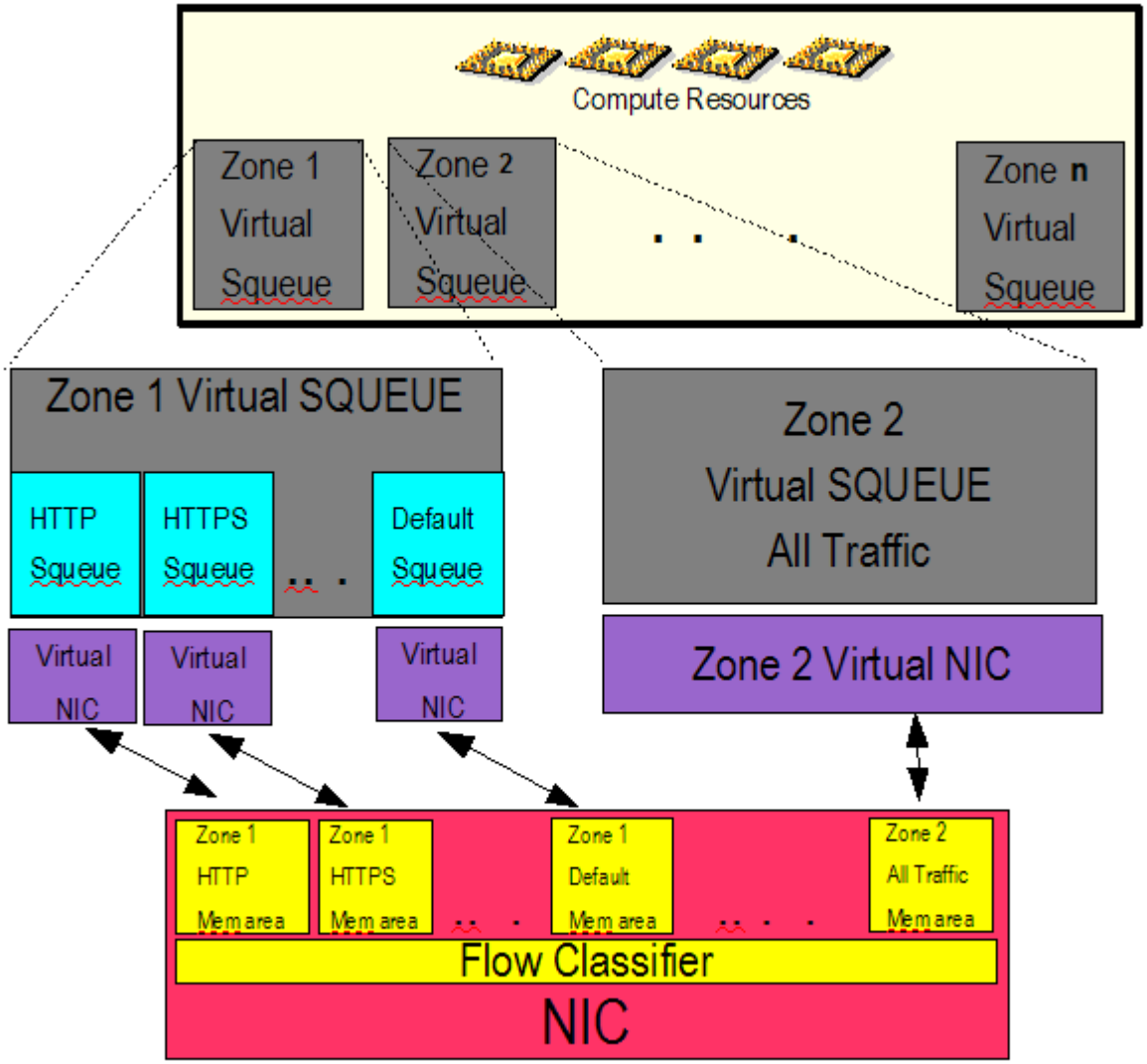
The VNICs are in the control path only. The data link layer is bypassed

The Squeue switches the MSI interrupt per stack between interrupt and polling mode and controls the rate of packet arrival for the Virtual stack

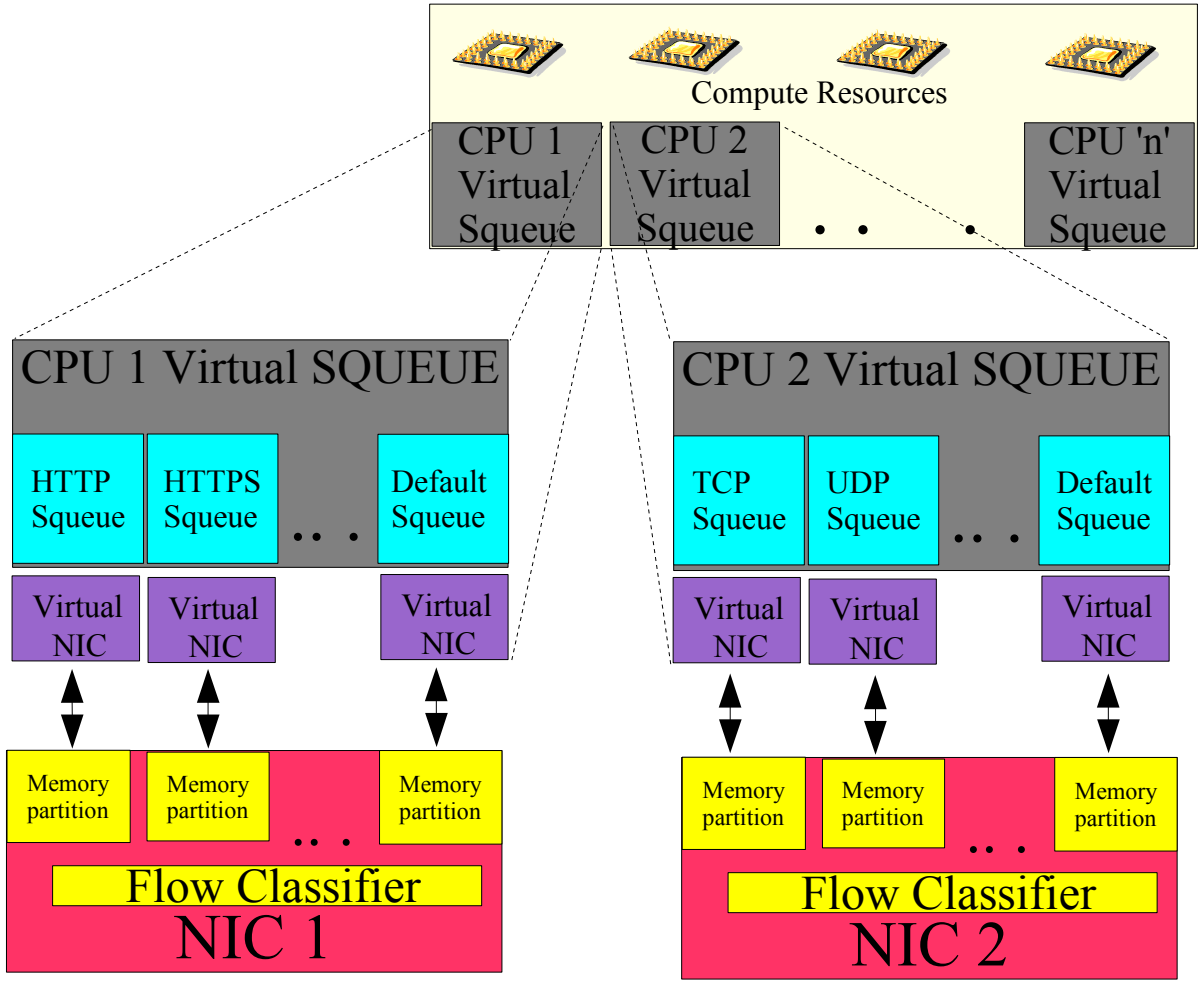
Virtual Stacks and containers

- Each Solaris container has its own virtual stack.
- When a container is created, the B/W, priority and number of possible virtual stacks within the container is specified.
- The Container administrator can configure the allocated virtual stacks to their own taste.
- Each Container can have its own routing table and firewall, and be tuned according to its own requirements.

Virtual Stack and Containers (Cont.)



Virtual Stacks – Services & Protocols



The VNICs are in the control path only. The data link layer is bypassed.

The Squeue switches the MSI interrupt per stack between interrupt and polling mode and controls the rate of packet arrival for the virtual stack

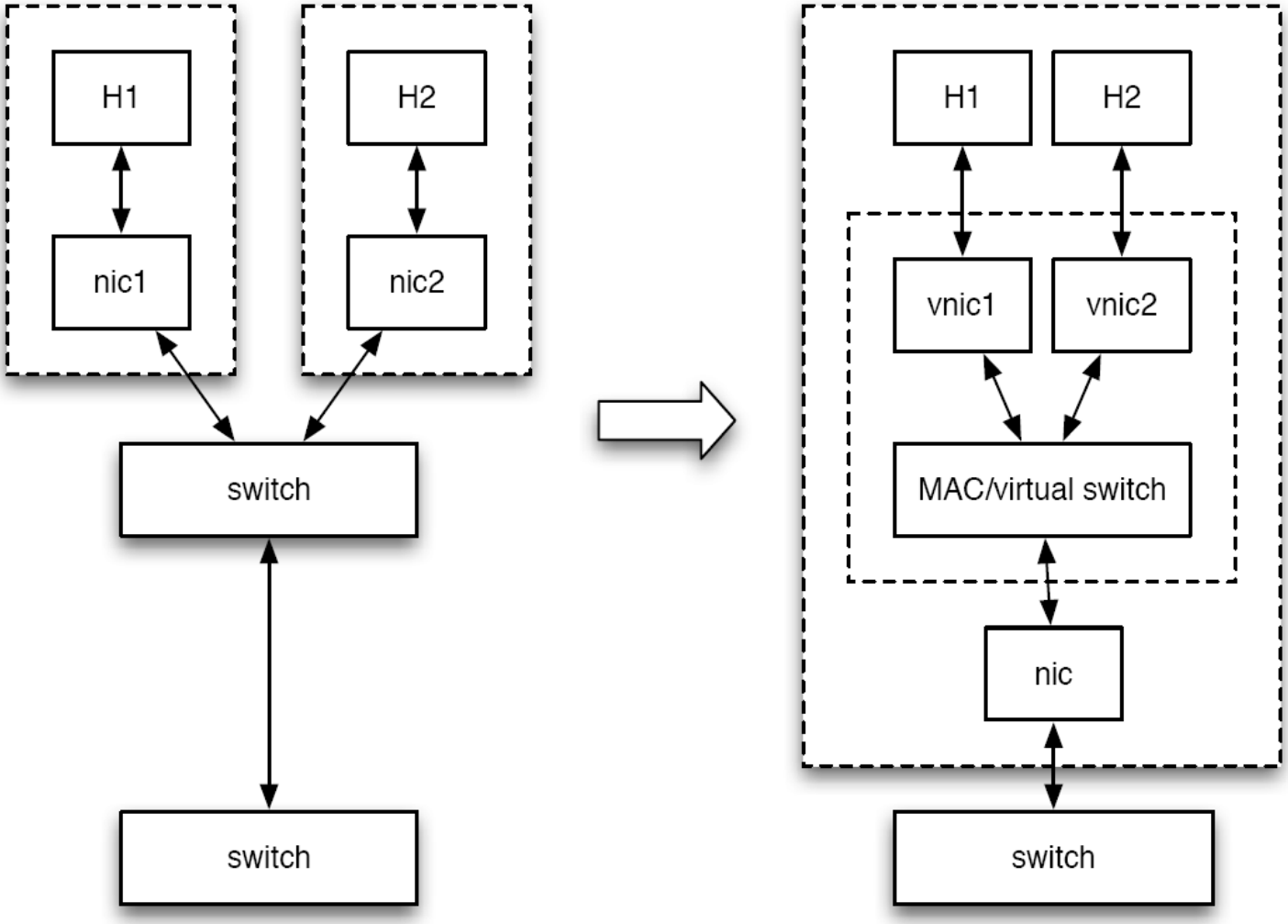
Soft vs Hard Virtualization

- Crossbow is evolving Solaris soft virtualization strategy
  - Containers provide the virtual application environment
  - Crossbow virtual stacks associated with Containers and CPU resource pool allow vertical partitioning on the machine
- Crossbow is complementary to Hard virtualization and allow network resource control for other virtual machines

New Concepts with Crossbow

- VNIC
  - Virtual NICs created on top of physical NIC. These appear as physical NIC to rest of the system
- Anchor NIC or Etherstub
  - They are Virtual stubs on top of which VNICs can be created. This is useful when testing a virtual private network inside the system.
- Virtual switch
  - MAC clients sharing the same underlying NIC are connected via a virtual switching mechanism.

Virtual Switching



Crossbow: Administrative Interface

- dladm(1M)
  - Manage creation, deletion and setting properties of VNICs
- netrcm(1M)
  - Set the bandwidth related attributes for any service, protocol or virtual machine
- ifconfig(1M)
  - Plumbing Virtual NICs
- flowadm(1M)
  - Managing the flow control
- acctadm(1M)
  - Accounting Tasks

Administrative Interface (Contd...)

- Alternatively, *cfg commands for virtual machines can be modified to take B/W, pri, phys/virt interfaces, IP addresses etc
  - `zonecfg -z new_zone`
  - `zonecfg:new_zone> create`
  - `zonecfg:new_zone> net phys=bge1`
  - `zonecfg:new_zone> net virt=eth0`
  - `zonecfg:new_zone> net bw=30Mbps`
  - `zonecfg:new_zone> net pri=hi`
  - `zonecfg:new_zone> net ip_addr=a.b.c.d`
- Similar mechanism for Xen/ldom etc
- Within a virtual machine, local admin can use dladm or netrcm to create more VNICs or policies

Accounting & History

- Finer grain accounting comes for free
- We can now do per squeue accounting to track usage by a container, service or protocol
- A userland daemon can pull the statistics out at fixed interval and do accounting etc.
- Usage history available for for all NICs, VNICs, and flows

Defense against DOS/DDOS

- DDOS have the ability to cripple entire server forms and all services offered by them
- Only the impacted services or virtual machine takes the hit instead of the entire grid
- Under attack, impacted services start all new connections under lower priority virtual stack with limited bandwidth
- Connections transition to appropriate priority stacks after application authentication

Summary

- Leading edge Virtualization technology
- Creates new paradigms to consolidate and deploy virtual network devices
- Provide policies to support class of services that span a large collection of machines

Unlocks new opportunities

Join Us...

- Our communities and projects are open on OpenSolaris.org:
  - Networking: http://opensolaris.org/os/community/networking
  - CrossBow: http://opensolaris.org/os/project/crossbow
- Where you will find:
  - Lively discussions, design docs, FAQs, source code drops, preliminary binary releases, etc...

Bibliography

- http://www.opensolaris.org/os/project/crossbow
- http://blogs.sun.com/sunay/
- http://blogs.sun.com/iyer/
- http://blogs.sun.com/droux/

THANK YOU!

CrossBow:
Network Virtualization Tutorial

Vineeth Pillai
SUN Microsystems Prague

Vineeth.pillai@sun.com